1. Suppose that in a batch of 20 items, 3 are defective. If 5 of the items are sampled at random:

(a) What is the probability that none of the sampled items are defective?

$X \sim \text{Hypergeometric}(5, 3, 20)$

$$P(X=0) = \frac{\binom{3}{0}\binom{17}{5}}{\binom{20}{5}} = \frac{91}{228} \approx 0.399$$

R: dhyper(0, 3, 17, 5)

Mathematica: PDF[HypergeometricDistribution[5, 3, 20], 0] // N

(b) What is the probability that exactly 1 of the sampled items are defective?

$$P(X=1) = \frac{\binom{3}{1}\binom{17}{4}}{\binom{20}{5}} = \frac{35}{76} \approx 0.461$$

R: dhyper(1, 3, 17, 5)

Mathematica: PDF[HypergeometricDistribution[5, 3, 20], 1] // N

(c) What is the probability that exactly 4 of the sampled items are defective?

Since the population contains only 3 defective items:

$$P(X=4) = 0 = \frac{\binom{3}{4}\binom{17}{1}}{\binom{20}{5}}$$

This is defined to be zero.

(d) On average how many defective items will be found in a random sample of 5 items?

Find the mean of X:

$$E(X) = n \cdot \frac{M}{N} = 5 \cdot \frac{3}{20} = \frac{3}{4}$$

(e) What is the probability that the number of defective items sampled is within 2 standard deviations of the mean number?

Find the standard deviation of X:

$$\text{Var}(X) = n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = 5 \cdot \frac{3}{20}\left(1 - \frac{3}{20}\right)\left(\frac{15}{19}\right) = \frac{153}{304} \approx 0.503$$

$$\sigma_X = \sqrt{\frac{153}{304}} \approx 0.709$$

Now find the requested probability:

$$P(|X - \mu| < 2\sigma_X) = P(-0.668 < X < 2.168)$$
$$= P(X \leq 2)$$

since X takes only nonnegative integer values

$$= \frac{113}{114} \approx 0.991$$

R: phyper(2, 3, 17, 5)

Mathematica: CDF[HypergeometricDistribution[3, 5, 20], 2]

2. Let X be a hypergeometric random variable with parameters n, M, and N. Let Y be a binomial random variable with parameters n and $p = \frac{M}{N}$. How does $E(X)$ compare to $E(Y)$? How does $\mathrm{Var}(X)$ compare to $\mathrm{Var}(Y)$?

The means of X and Y are equal:

$$E(X) = n \cdot \frac{M}{N} = n \cdot p = E(Y)$$

The variance of X is less than the variance of Y if $n > 1$:

$$\mathrm{Var}(X) = n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\underbrace{\left(\frac{N-n}{N-1}\right)}_{\leq 1} \leq np(1-p) = \mathrm{Var}(Y)$$

So $\mathrm{Var}(X) < \mathrm{Var}(Y)$ if $n > 1$.

(If $n=1$, both X and Y have the same Bernoulli distribution.)

3. An unknown number, N, of animals inhabit a certain region. To estimate the size of the population, ecologists perform the following experiment: They first catch M of these animals, mark them in some way, and release them. After allowing the animals to disperse throughout the region, they catch n of the animals and count the number, X, of marked animals in this second catch.

The ecologists want to make a *maximum likelihood estimate* of the population size N. This means that if the observed value of X is x, then they estimate the population size to be the integer N that maximizes the probability that $X = x$. Help them complete this estimate as follows.

(a) What assumptions are necessary to say that X has a hypergeometric distribution?

Assume that the population of animals remains fixed between the two catches, and that each animal is equally likely to be caught.

(b) Let $P_x(N)$ be the probability that $X = x$ given that $X \sim \text{Hypergeometric}(n, M, N)$. Write down a formula for $P_x(N)$

assume n, M, and N are known

$$P_x(N) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

(c) Simplify the ratio $\frac{P_x(N)}{P_x(N-1)}$. *Hint*: use `FullSimplify` in Mathematica!

N is a built-in function in Mathematica, so here we use **pop** for the population size:

```
In[18]:= FullSimplify[(Binomial[M, x] Binomial[pop - M, n - x])/Binomial[pop, n] / (Binomial[M, x] Binomial[pop - 1 - M, n - x])/Binomial[pop - 1, n]]

Out[18]= ((-M + pop) (-n + pop))/(pop (-M - n + pop + x))
```

This is $P_x(N)$. This is $P_x(N-1)$.

Thus: $$\frac{P_x(N)}{P_x(N-1)} = \frac{(N-M)(N-n)}{N(N+x-M-n)}$$ Note that $P_x(N) \geq P_x(N-1)$ if and only if this fraction is ≥ 1.

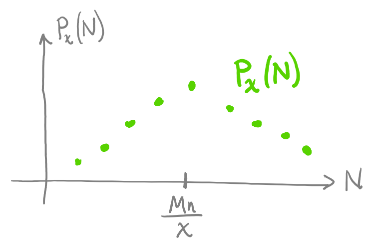
(d) Show that $\frac{P_x(N)}{P_x(N-1)} \geq 1$ if and only if $N \leq \frac{Mn}{x}$.

$$\frac{P_x(N)}{P_x(N-1)} \geq 1 \quad \text{iff} \quad \frac{(N-M)(N-n)}{N(N+x-M-n)} \geq 1$$

$$\text{iff} \quad N^2 - Nn - MN + Mn \geq N^2 + Nx - MN - Nn$$

$$\text{iff} \quad Mn \geq Nx$$

$$\text{iff} \quad \frac{Mn}{x} \geq N$$



(e) Explain why $P_x(N)$ attains its maximum value when N is the largest integer less than or equal to $\frac{Mn}{x}$. What is the most likely population size N?

If $N \leq \frac{Mn}{x}$, then $P_x(N) \geq P_x(N-1)$, so population size N is more likely than population size N-1.

However, if $N > \frac{Mn}{x}$, then $P_x(N) < P(N-1)$, so population size N is less likely than population size N-1.

Thus, the most likely population size is the largest integer N that is less than or equal to $\frac{Mn}{x}$.

(f) If $\underbrace{M = 30}_{\text{marked}}$, $\underbrace{n = 20}_{\text{captured}}$, and $\underbrace{x = 7}_{\text{captured animals that are marked}}$, what is the maximum likelihood estimate for N?

$$\frac{Mn}{x} = \frac{30(20)}{7} \approx 85.7,$$ so the estimate is $N = 85$.

4. Urn 1 contains 100 balls, 10 of which are red. Let X_1 be the number of red balls in a random sample of size 50 from Urn 1. Urn 2 contains 100 balls, 50 of which are red. Let X_2 be the number of red balls in a random sample of size 10 from Urn 2.

(a) Use technology to compute the pmf of X_1. Display the values as a table. Then do the same for the pmf of X_2. What do you notice?

Using Mathematica:

```
In[9]:= Table[PDF[HypergeometricDistribution[50, 10, 100], x], {x, 0, 10}] // N
Out[9]= {0.00059342, 0.00723683, 0.0379933, 0.113096, 0.211413,
  0.259334, 0.211413, 0.113096, 0.0379933, 0.00723683, 0.00059342}
```

] pmf of X_1

```
In[11]:= Table[PDF[HypergeometricDistribution[10, 50, 100], x], {x, 0, 10}] // N
Out[11]= {0.00059342, 0.00723683, 0.0379933, 0.113096, 0.211413,
  0.259334, 0.211413, 0.113096, 0.0379933, 0.00723683, 0.00059342}
```

] pmf of X_2

The two pmfs are the same!

(b) Change the numbers 100, 10, and 50 in this problem and recompute the pmfs of X_1 and X_2. What do you notice?

For example, if $X_1 \sim \text{Hypergeometric}(45, 12, 80)$
and $X_2 \sim \text{Hypergeometric}(12, 45, 80)$
then X_1 and X_2 again have the same pmf.

(c) Make a conjecture about when two hypergeometric random variables have the same pmf.

If $X_1 \sim \text{Hypergeometric}(a, b, N)$ and $X_2 \sim \text{Hypergeometric}(b, a, N)$,
then X_1 and X_2 have the same pmf.